Variance Modeling Research for the Small Area Health Insurance Estimates Program [1]

Mark Bauder and Samuel Szelepka

U.S. Census Bureau
Washington, D.C.

Proceedings of the 2013 Federal Committee on Statistical Methodology (FCSM) Research Conference

1 Introduction

The U.S. Census Bureau's Small Area Health Insurance Estimates Program (SAHIE) produces estimates of the numbers and proportions with and without health insurance by income and demographic domains within states or counties. SAHIE uses estimates from the American Community Survey (ACS) as data in the models. In 2012, the National Opinion Research Center (NORC) reviewed the SAHIE methodology. One of NORC's recommendations was that SAHIE investigate the use of a generalized variance function (gvf) to estimate the ACS survey variances outside of the estimation of the rest of the model. This paper reports results of that investigation.

2 The SAHIE model

SAHIE produces estimates of the numbers by income and demographic domains within states or counties. For states, the demographic groups are defined by age, race, sex and income group; for counties, they are defined by age, sex and income group. The income groups are defined by the "income to poverty ratio" (IPR), which is the ratio of family income to the federal poverty threshold appropriate for that family. We fit models separately for states and counties except that county estimates are controlled to state estimates that are themselves controlled to national level estimates from the American Community Survey (ACS).

The SAHIE model consists of two largely distinct submodels, one for estimating proportions in IPR categories within state/age/race/sex or county/age/sex groups, and one for estimating proportions with health insurance within state/age/race/sex/IPR or county/age/sex/IPR categories. In each part of the model, survey data and administrative data are modeled, conditional on the true proportions. The base level of modeling includes the full cross-classification of: five age groups, four race groups (for states), two sexes, and five IPR categories.[2]

SAHIE uses ACS estimates as data in the model. We model the ACS estimates of proportions in income groups, and of proportions with health insurance within those income groups. The current model for the ACS estimates is as follows. Let a index a state/age/race/sex or county/age/sex group, and let i index an IPR category. Let S_a be sample size in the a^{th} state/age/race/sex group.[3] Let $\hat{p}_{ai}^{\text{IPR}}$ be the ACS estimate of the proportion, p_{ai}^{IPR}, in income category i within group a. The model for $\hat{p}_{ai}^{\text{IPR}}$, conditional on p_{ai}^{IPR} and

[1] This report is released to inform interested parties of ongoing research and to encourage discussion of work in progress. Any views expressed on statistical, methodological, technical, or operational issues are those of the authors and not necessarily those of the U.S. Census Bureau.

[2] Many of the estimates we release are aggregates over these base-level classifications.

[3] Here and elsewhere, we use 'sample size' to refer to an unweighted ACS person count.

parameters, is given by

$$\big(\hat{p}^{\text{IPR}}_{a1}, \ldots, \hat{p}^{\text{IPR}}_{a4}\big)' \mid p^{\text{IPR}}_{a1}, \ldots, p^{\text{IPR}}_{a4}, \lambda_0, \lambda_1 \tag{1}$$

$$\sim \mathcal{N}((p^{\text{IPR}}_{a1}, \ldots, p^{\text{IPR}}_{a4})',\ \Sigma^{\text{IPR}}_{a}) \tag{2}$$

$$(\Sigma^{\text{IPR}}_{a})_{ii} = \frac{\lambda_0 p^{\text{IPR}}_{ai}(1 - p^{\text{IPR}}_{ai})}{S_a^{\lambda_1}}\ ,\ \ i = 1, \ldots, 4 \tag{3}$$

$$(\Sigma^{\text{IPR}}_{a})_{ij} = \rho_{aij}\sqrt{(\Sigma^{\text{IPR}}_{a})_{ii}\,(\Sigma^{\text{IPR}}_{a})_{jj}}\ ,\ \ i \neq j \tag{4}$$

$$\text{where}\quad \rho_{aij} = -\sqrt{\frac{p^{\text{IPR}}_{ai}p^{\text{IPR}}_{aj}}{(1 - p^{\text{IPR}}_{ai})(1 - p^{\text{IPR}}_{aj})}}\ . \tag{5}$$

Note that in (3) the survey sampling variance is modeled, and the parameters λ_0 and λ_1 are estimated jointly with the rest of the unknowns in the SAHIE model. In addition, (5) expresses the assumption that the correlations among the survey estimates $\hat{p}^{\text{IPR}}_{a1}, \ldots, \hat{p}^{\text{IPR}}_{a4}$ are those of a multinomial.

Let $\hat{p}^{\text{IC}}_{ai}$ be the ACS estimate of p^{IC}_{ai}, the proportion insured in the i^{th} IPR category within the state/age/race/sex or county/age/sex group a. Let S_{ai} be the sample size in group a, IPR category i. The model for $\hat{p}^{\text{IC}}_{ai}$, conditional on p^{IC}_{ai} and parameters, is as follows:

$$\hat{p}^{\text{IC}}_{ai} | \, p^{\text{IC}}_{ai}, \lambda_0, \lambda_1, \zeta_0, \zeta_1 \begin{cases} = 0 & \text{with probability } p^{(0)}_{ai} \\ = 1 & \text{with probability } p^{(1)}_{ai} \\ \sim \text{Beta}(\mathfrak{a}_{ai}, \mathfrak{b}_{ai}) & \text{with probability } 1 - p^{(0)}_{ai} - p^{(1)}_{ai} \end{cases} \tag{6}$$

with

$$\mathbb{E}(\hat{p}^{\text{IC}}_{ai} | \, p^{\text{IC}}_{ai}) = p^{\text{IC}}_{ai} \tag{7}$$

$$\text{var}(\hat{p}^{\text{IC}}_{ai} | \, p^{\text{IC}}_{ai}, \lambda_0, \lambda_1) = \frac{\lambda_0 p^{\text{IC}}_{ai}(1 - p^{\text{IC}}_{ai})}{S_{ai}^{\lambda_1}} \tag{8}$$

$$p^{(0)}_{ai} = (1 - p^{\text{IC}}_{ai})^{1+\zeta_0(S_{ai}-1)} \tag{9}$$

$$p^{(1)}_{ai} = (p^{\text{IC}}_{ai})^{1+\zeta_1(S_{ai}-1)}\ . \tag{10}$$

The parameters $\lambda_0, \lambda_1, \zeta_0$ and ζ_1 are unknown and allowed to vary by demographic group. The parameters $\mathfrak{a}_{ai}$ and $\mathfrak{b}_{ai}$ are functions of the mean and variance of $\hat{p}^{\text{IC}}_{ai}$ together with $p^{(0)}_{ai}$ and $p^{(1)}_{ai}$. In equation (8), the ACS sampling variance is modeled jointly with other unknowns in the SAHIE model.

The alternative we consider here is to replace the model for $(\Sigma^{\text{IPR}}_{a})_{ii}$ in (3) and the model for $\text{var}(\hat{p}^{\text{IC}}_{ai})$ in (8) with plugged-in estimates from a gvf.

3 Variance modeling

3.1 Design effects

It is useful to think of sampling variances of survey estimates in terms of *design effects.* We define the design effect of a survey estimator $\hat{x}$ to be

$$\text{deff}(\hat{x}) = \frac{\text{variance of } \hat{x}}{\text{variance of } \hat{x} \text{ under simple random sampling}}.$$

In the present paper, we are concerned with ACS estimates of proportions. We define the design effect of an estimator, $\hat{p}$, of a proportion p to be

$$\text{deff}(\hat{p}) = \frac{\text{var}(\hat{p})}{p(1-p)/S} \tag{11}$$

where S is the relevant sample size. For our purposes, S is the unweighted count of people in sample in the domain that constitutes the denominator of the rate.

For our analyses, we use estimates of the design effect:

$$\widehat{\text{deff}}(\hat{p}) = \frac{\widehat{\text{var}}(\hat{p})}{\hat{p}(1-\hat{p})/S} \; . \tag{12}$$

For the research here, we model the variances by modeling the design effects. The models are of the form

$$\log(\widehat{\text{deff}}) = State_s + (z_{ai})^T \gamma + \epsilon_{ai} \tag{13}$$

where the state effects, $State_s$, might not be included, and z_{ai} may contain predictors in addition to indicators of age, race, sex and income category and their interactions. We then assume the sampling variance to be

$$\text{var}(\hat{p}_{ai}) = \frac{p_{ai}(1-p_{ai})}{S_{ai}} \text{deff}_{ai} \; . \tag{14}$$

If z_{ai} does not contain the sample size, then the variance varies inversely with the sample size. If z_{ai} does not include p_{ai} then the variance is proportional to $p_{ai}(1-p_{ai})$.

3.2 ACS variance estimation

We use estimates of ACS variances. The ACS employs the successive differences replication method (U.S. Census Bureau (U.S. Census Bureau (2009)), Chapter 12; Judkins (Judkins (1990))) to produce estimates of the variances of its estimates. The method involves 80 replicates.

4 Analysis of design effects for $\hat{p}^{\text{IC}}$

In this section we report results of exploratory analyses of design effects. We begin by considering the variances of $\hat{p}^{\text{IC}}_{ai}$, the estimates of the proportion insured. Figure 1 contains boxplots for states and counties of estimated design effects against sextile of sample size, with separate plots by income category and by age group.[4] For consistency among plots, we take sextiles over sample sizes for all observations, rather than within the income and age categories. As a result, the boxes do not generally represent the same number of points.

From Figure 1, it appears that the design effects increase with increasing sample size. Thus it appears that the variances decrease at a rate of less than the inverse of the sample size. There does not appear to be much difference between IPR categories.

The plots for states and counties look different, especially in the highest two IPR categories. However, note that the boxes represent different sample sizes between states and counties. Some, but not all of the difference in appearance between the state and county plots may be due to the fact that the first box for states contains the sample sizes of several boxes for counties. Increases in design effects for the larger sample sizes that can be seen in the plots for states cannot be seen in the plots for counties, because there are fewer large sample sizes for counties.

Figure 2 contains similar plots by age. The variances for the lowest age group behave differently from those for the other age groups. This is likely due to clustering. Within housing units, income level and insured

[4] Here, and in all plots, estimated design effects are bottom- and top- coded at 0.25 and 4 respectively.

status are highly correlated within age/sex groups. It is relatively rare that two adults in the same housing unit are in the same age and sex group. It is less rare, in households with children, that two children of the same sex are in the same household. In the ACS sample for 2010, among housing units in which there is at least one male age 0 to 17, there is an average of 1.4 males age 0 to 17 in the housing unit. The average is 1.4 for females age 0 to 17 as well. In each other age/sex group used in SAHIE, in housing units in which there is at least one person in the age/sex group, the mean number in that age/sex group per housing unit is 1.0 for males or 1.1 for females. The numbers are similar when broken into age/race/sex groups.

Figure 3 contains plots similar to those in Figure 2, but by race. The variances do not appear to behave much differently by race.

4.1 Behavior of variance estimates for different ranges of proportions

Figure 4 contains boxplots of design effects against sextile of sample size, with separate plots by sextile of $\hat{p}^{\text{IC}}$. We note that the dependence of the estimated design effects on sample size becomes much more pronounced as the $\hat{p}^{\text{IC}}$'s approach 1.0. In addition, for $\hat{p}^{\text{IC}}$ intervals closer to one, the design effects appear unreasonable at low sample sizes. For states, when $\hat{p}^{\text{IC}}$ is larger than 0.94, the large majority of design effect estimates in the first sextile of sample sizes, containing sample sizes between two and 17,[5] are smaller than 1.0, and the majority of design effect estimates are less than 1.0 in the second sextile, containing sample sizes 18 to 43. Even in the sextile of $\hat{p}^{\text{IC}}$ covering 0.88 to 0.94, the majority of design effects are smaller than 1.0 for the smallest sextile of sample sizes. In these two intervals of $\hat{p}^{\text{IC}}$, the design effect variances approach 2.0 for large sample sizes. For $\hat{p}^{\text{IC}}$ smaller than about 0.7, the design effect estimates show little dependence on sample size. Design effect estimates behave similarly for counties.

In Figure 5, we look at the phenomenon of the previous paragraph in another way. The figure contains boxplots against sextile of $\hat{p}^{\text{IC}}$, with separate plots by sample size sextile. We see that for small sample sizes, the design effect estimates get small for $\hat{p}^{\text{IC}}$ close to 1.0.

4.2 Discussion

Some of the results shown above indicate weaknesses in the variance estimation method, and suggest that we leave some observations out when using variance estimates in models aimed at predicting variances.

We expect design effects generally to be larger than 1.0. We expect clustering, nonresponse, and factors such as reweighting that go into creating estimates, generally to increase the variances of estimates from what they would have been under simple random sampling. Design effects could be less than one due to the fact that some totals are controlled. But we expect it to be rare that controls substantially reduce design effects for the small domains at which SAHIE estimates, especially where sample sizes are small.

Instead, we conclude from the plots in Figures 4 and 5 that for combinations small sample size and $\hat{p}^{\text{IC}}$ close to one, the variances are generally underestimated. Note that for largest sextiles of $\hat{p}^{\text{IC}}$, the variance estimates appear generally to be unreasonably low even for moderate sample sizes. For example, Figure 5 shows that for $\hat{p}^{\text{IC}}$ between 0.93 and 1.0, the median estimated design effect is smaller than 1.0 for observations with sample sizes between 25 to 51. Similarly, for small sample sizes, variance estimates appear generally unreasonable for $\hat{p}^{\text{IC}}$'s relatively far from 1.0, for example, in the second highest sextile of $\hat{p}^{\text{IC}}$'s.

Because the variance estimates often appear to be unreasonably small for small samples and for proportions close to one, it appears that a small unweighted count in the uninsured group is an indication that the variance estimate is likely to behave badly. That unweighted count is small for small samples, and also is small for larger samples when the estimated proportion is close to 1.0. Figure 6 contains plots by sextile of the unweighed uninsured count of the estimated design effects against sextiles of the estimated proportion insured. From these plots, it appears that the variance estimates behave reasonably when the unweighted count is around five or more.

[5] We leave observations with sample sizes of one out, because they have replicate variance estimates of zero.

Finally, we note that even when the sample sizes are large, there appears to be some dependence of design effect on the proportions. This suggests that the variance of $\hat{p}^{\text{IC}}$ is not proportional to $\hat{p}^{\text{IC}}(1-\hat{p}^{\text{IC}})$.

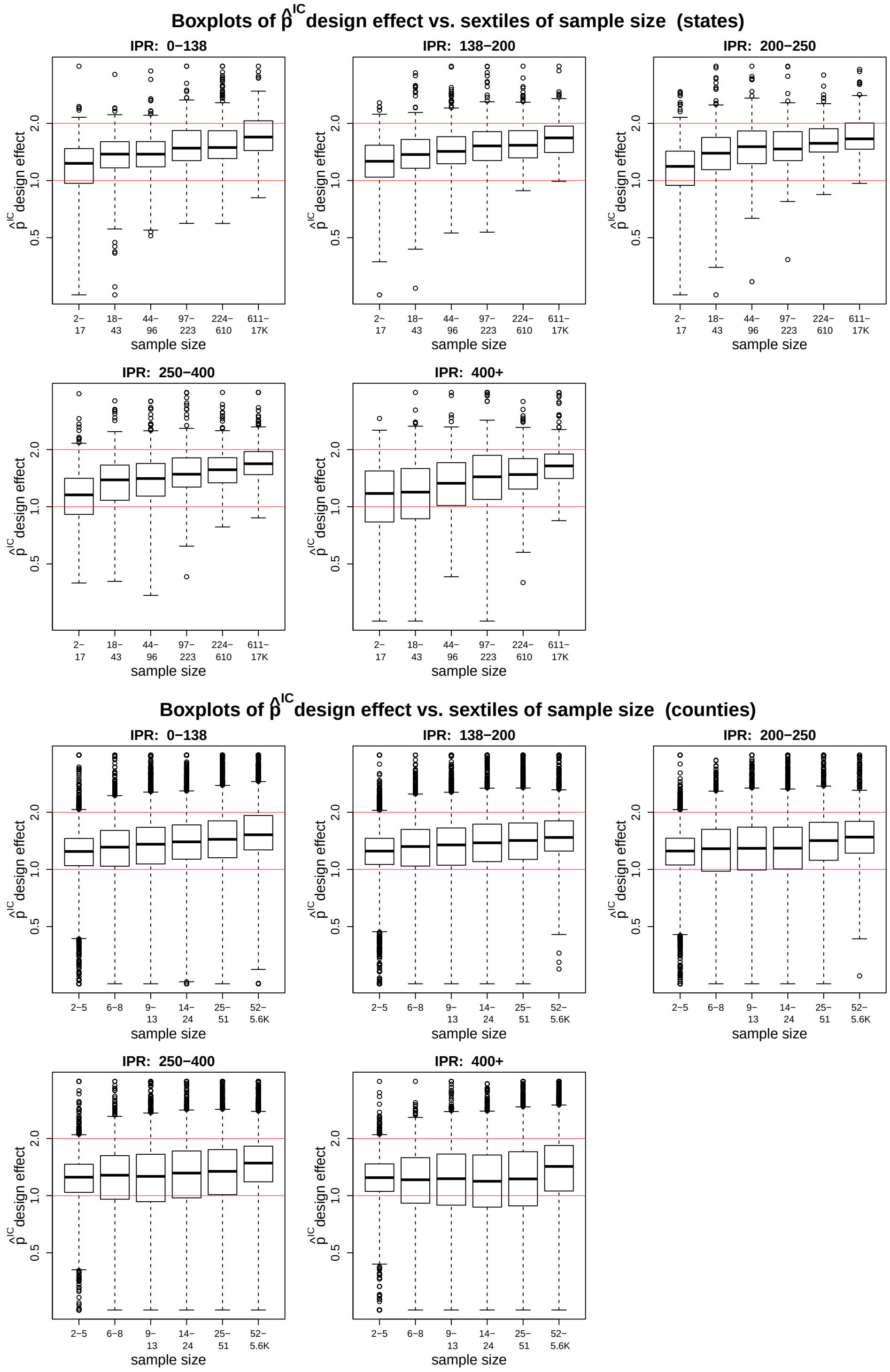

Figure 1. Boxplots of $\hat{p}^{\text{IC}}$ estimated design effects against sextile of sample size, by IPR category. For states (top) and counties (bottom). Here, and in all plots, boxes extend from the first to third quartiles, and design effect estimates are bottom- and top- coded at 0.25 and 4 respectively. Source: ACS 2010 unpublished estimates.

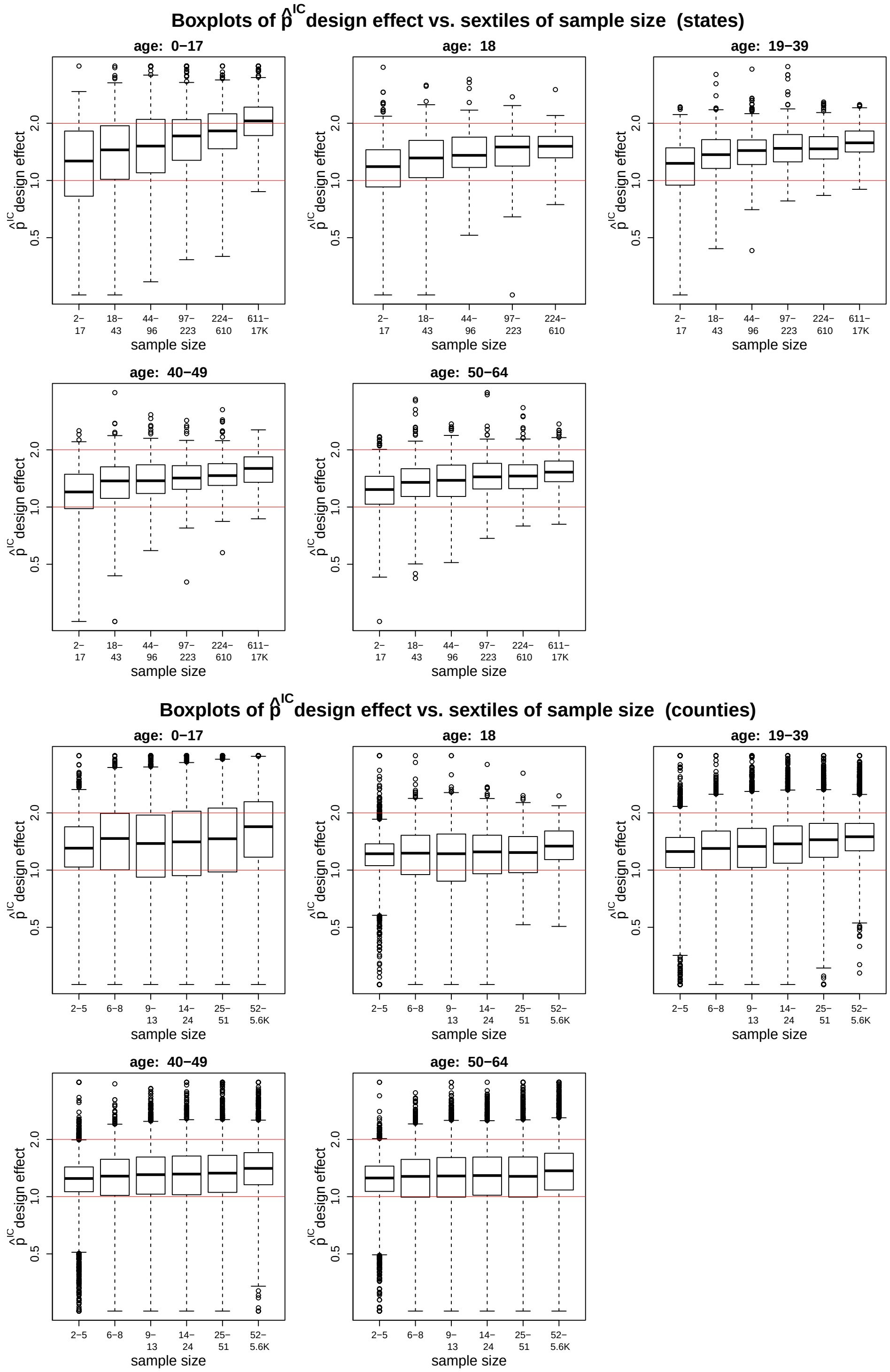

Figure 2. Boxplots of $\hat{p}^{\text{IC}}$ estimated design effects against sextile of sample size, by age category for states (top) and counties (bottom). Source: ACS 2010 unpublished estimates.

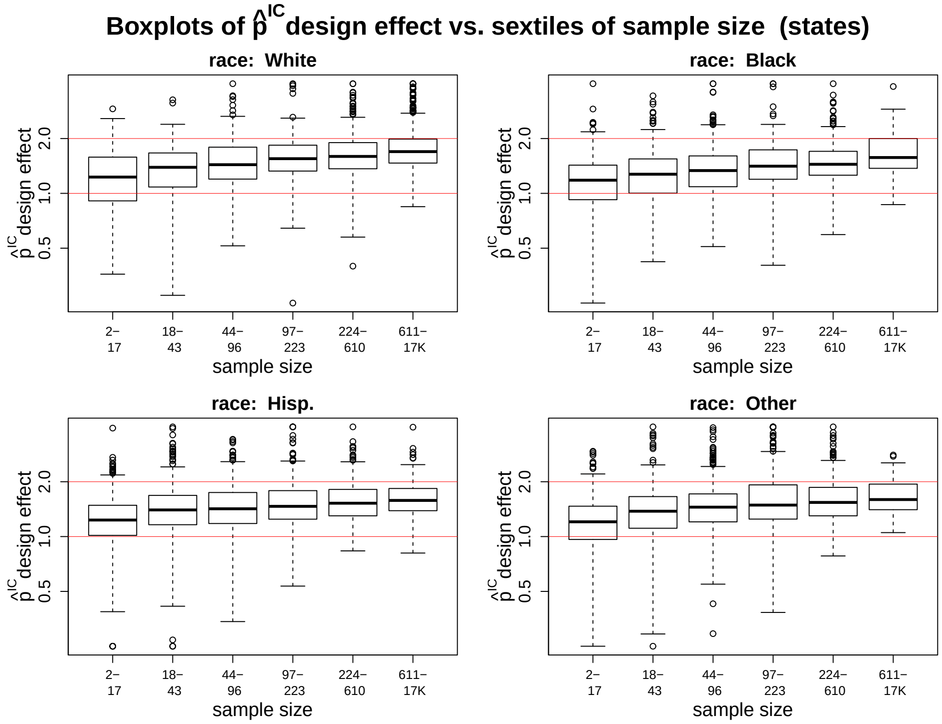

Figure 3. Boxplots of $\hat{p}^{\text{IC}}$ estimated design effects for state estimates against sextile of sample size, by race. Source: ACS 2010 unpublished estimates.

4.3 State effects

Figure 7 contains boxplots by state of design effects for $\hat{p}^{\text{IC}}$, for state estimates and for county estimates. The design effects appear to be relatively consistent by state. The design effects for county estimates appear different for the seventh least populated state. That state is Delaware and has only three counties. Thus its box is based on a small number of points, and its difference could be due to randomness.

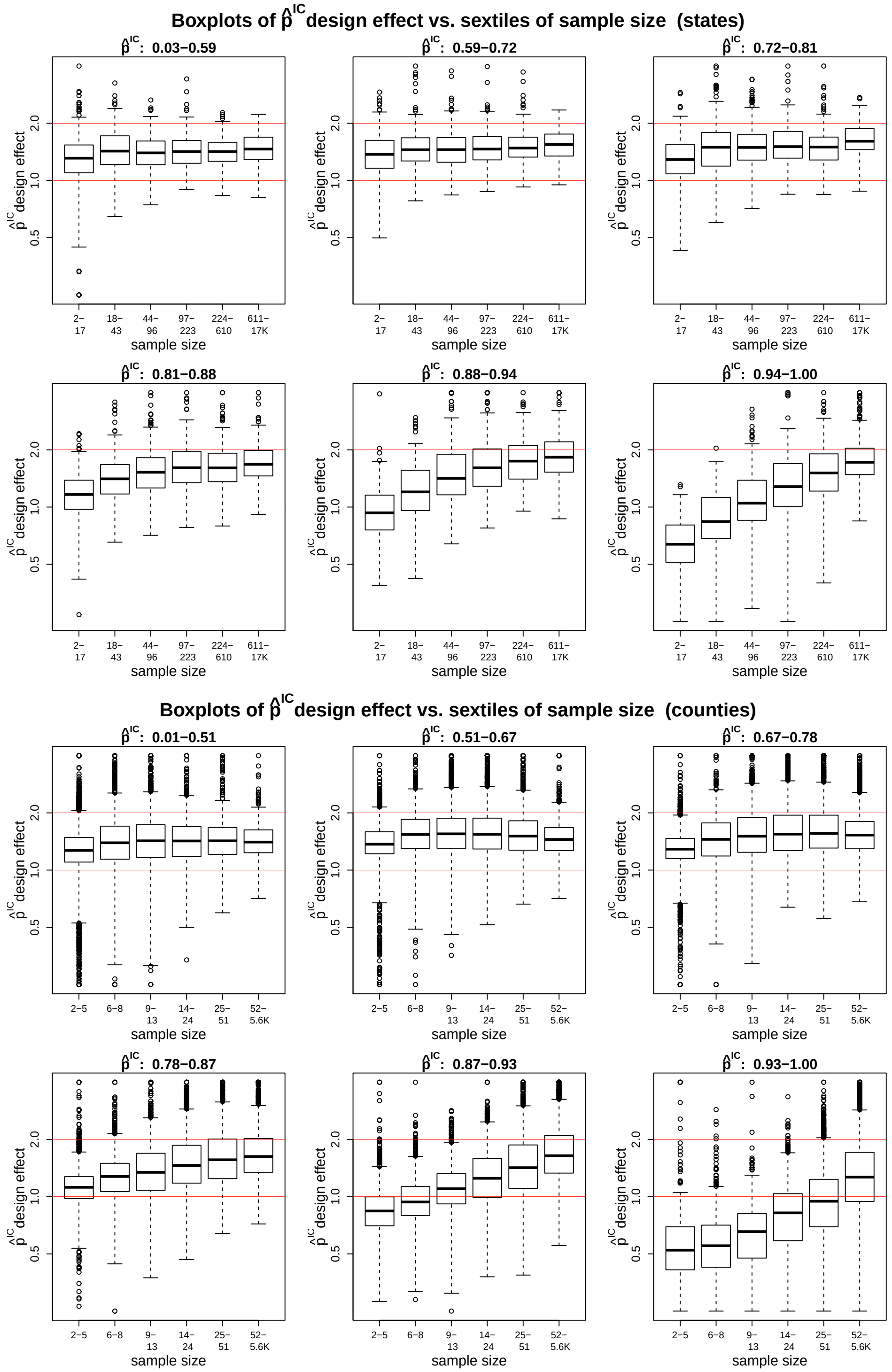

Figure 4. Boxplots of $\hat{p}^{\,\mathrm{IC}}$ estimated design effects against sextile of sample size, by sextile of $\hat{p}^{\,\mathrm{IC}}$. For states (top) and counties (bottom). Source: ACS 2010 unpublished estimates.

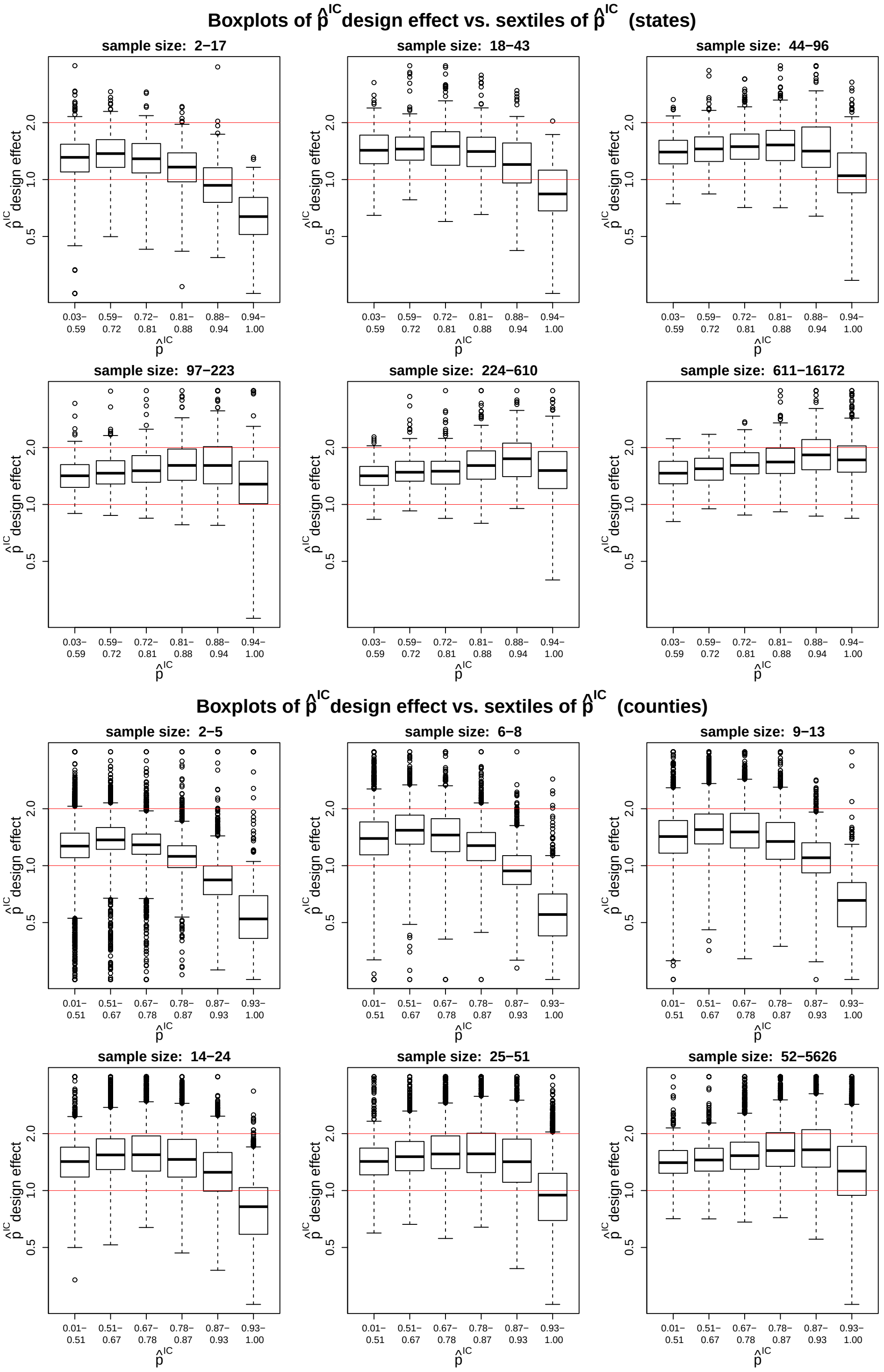

Figure 5. Boxplots of $\hat{p}^{\,\mathrm{IC}}$ estimated design effects against sextile of $\hat{p}^{\,\mathrm{IC}}$ by sextile of sample size. For states (top) and counties (bottom). Source: ACS 2010 unpublished estimates.

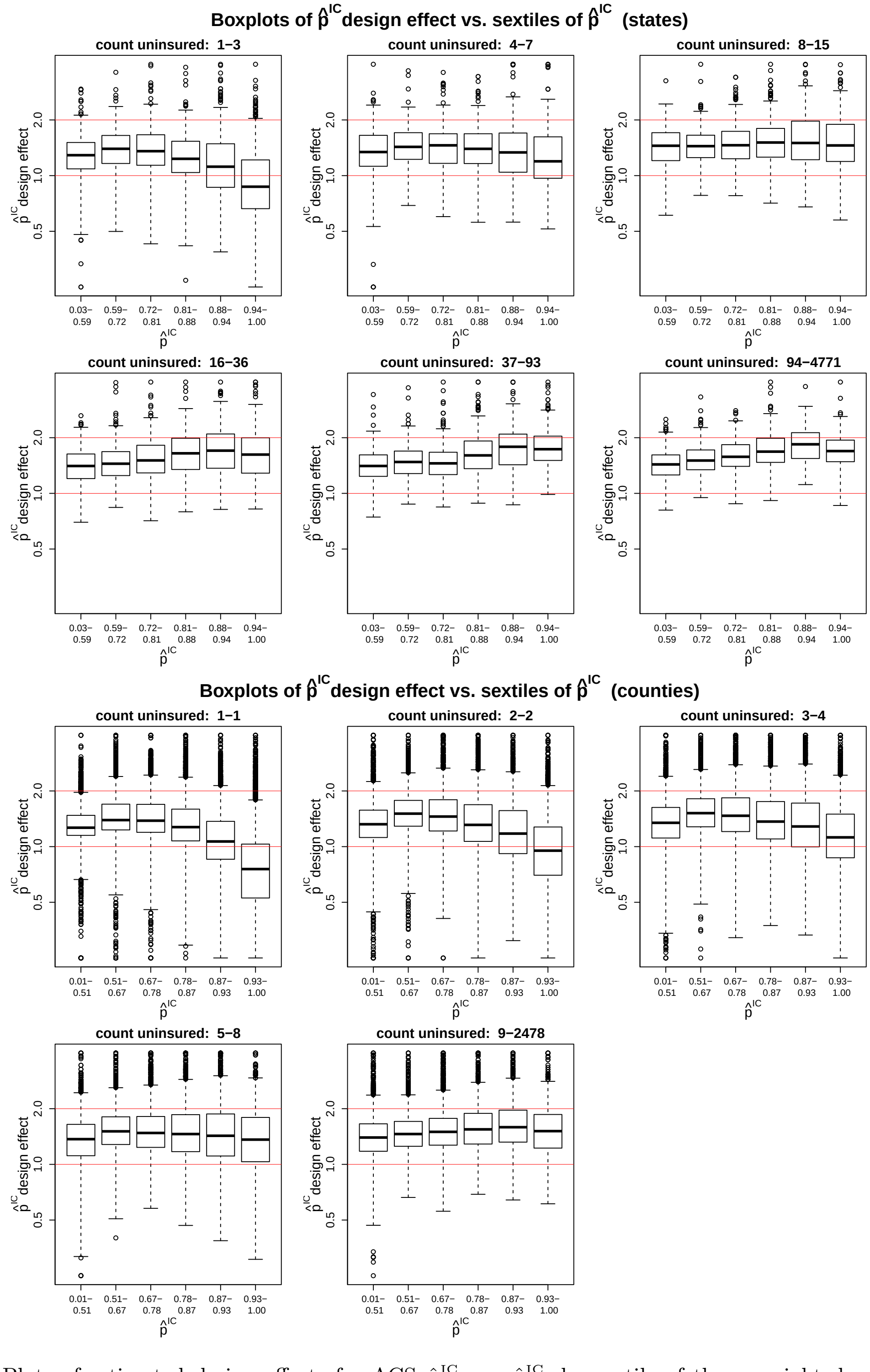

Figure 6. Plots of estimated design effects for ACS $\hat{p}^{\text{IC}}$ vs. $\hat{p}^{\text{IC}}$, by sextile of the unweighted count of number uninsured. For states (top) and counties (bottom). Source: ACS 2010 unpublished estimates.

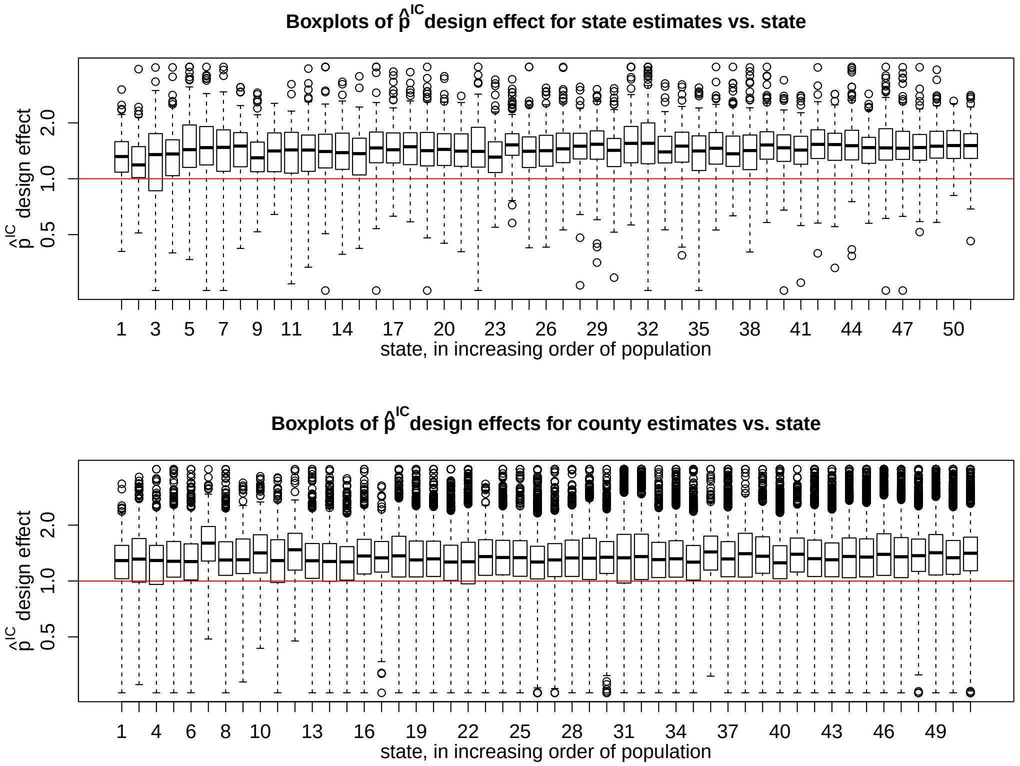

Figure 7. Boxplots of estimated design effects for ACS $\hat{p}^{\,\mathrm{IC}}$ vs. state, in increasing order of population. Top plot contains design effects for state/age/race/sex/IPR estimates. Bottom plot contains design effects for county/age/sex/IPR estimates. Source: ACS 2010 unpublished estimates.

5 Analysis of design effects for $\hat{p}^{\text{IPR}}$

In this section, we look at design effect estimates for $\hat{p}_{ai}^{\text{IPR}}$, the estimate of the proportion in the IPR category, i, among those in state/age/race/sex or county/age/sex group a. Recall that for a given a, $\sum_{i=1}^{5} \hat{p}_{ai}^{\text{IPR}} = 1$. Figure 8 contains boxplots by IPR category of $\hat{p}^{\text{IPR}}$ design effects, against sextile of state/age/race/sex or county/age/sex sample size. The patterns are similar to those for $\hat{p}^{\text{IC}}$. There appears to be some upward trend against sample size, suggesting that variances decrease less quickly than the inverse of the sample size. In Figures 9 and 10, we plot design effect estimates for states and counties against $\hat{p}^{\text{IPR}}$ for the five IPR categories for selected sample size groups. Note that these sample size groups are not quantiles. Unlike proportions insured, the proportions in the IPR categories are rarely near 1.0. However, they can be near zero. As the proportions get close to zero, we see a pattern similar to that for proportions insured near 1.0. For small samples, the design effects become unreasonably small.

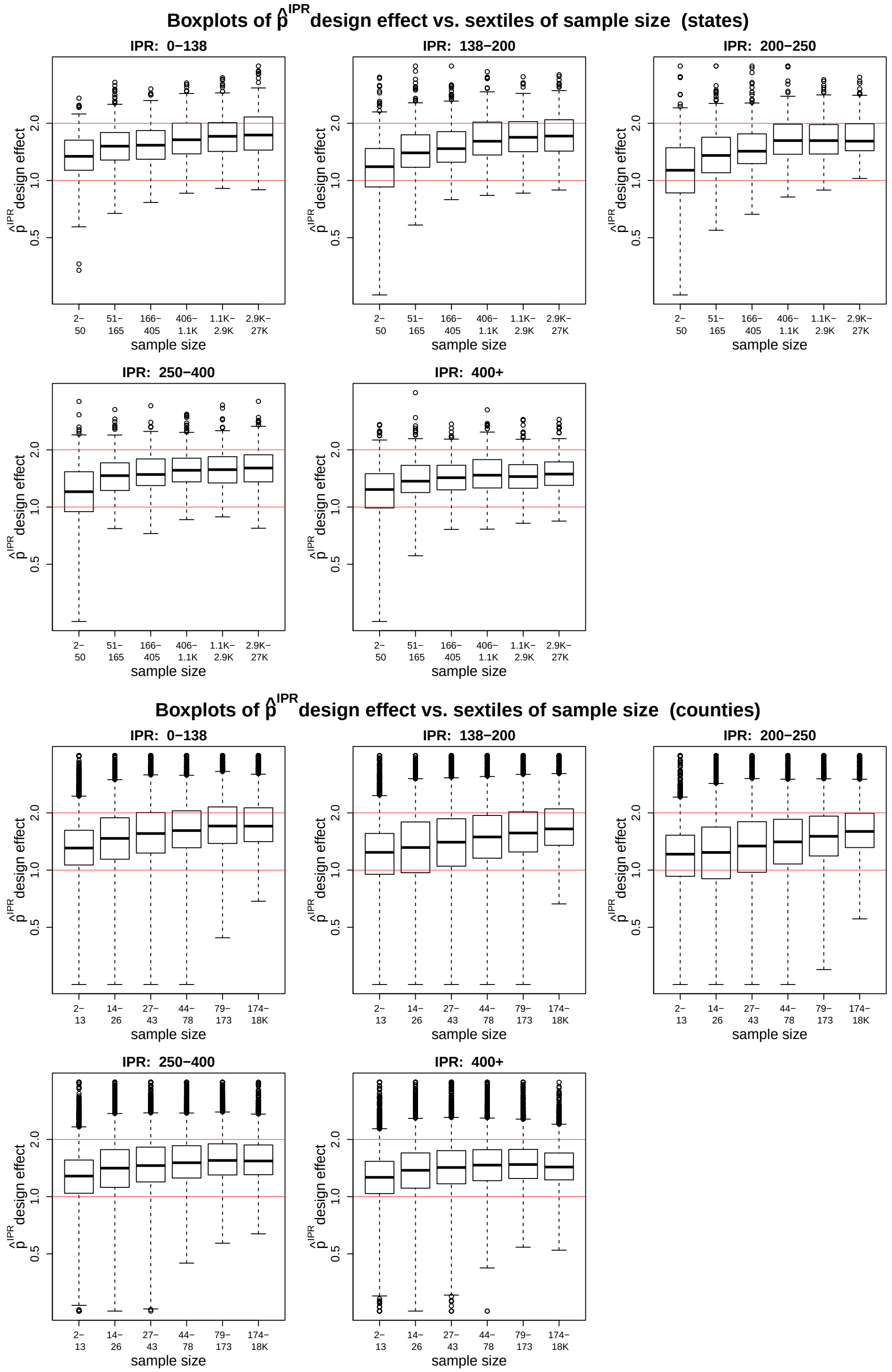

Figure 8. Boxplots of estimated design effects for ACS $\hat{p}^{\mathrm{IPR}}$ vs. sextile of sample size, by IPR category. For states (top) and counties (bottom). Source: ACS 2010 unpublished estimates.

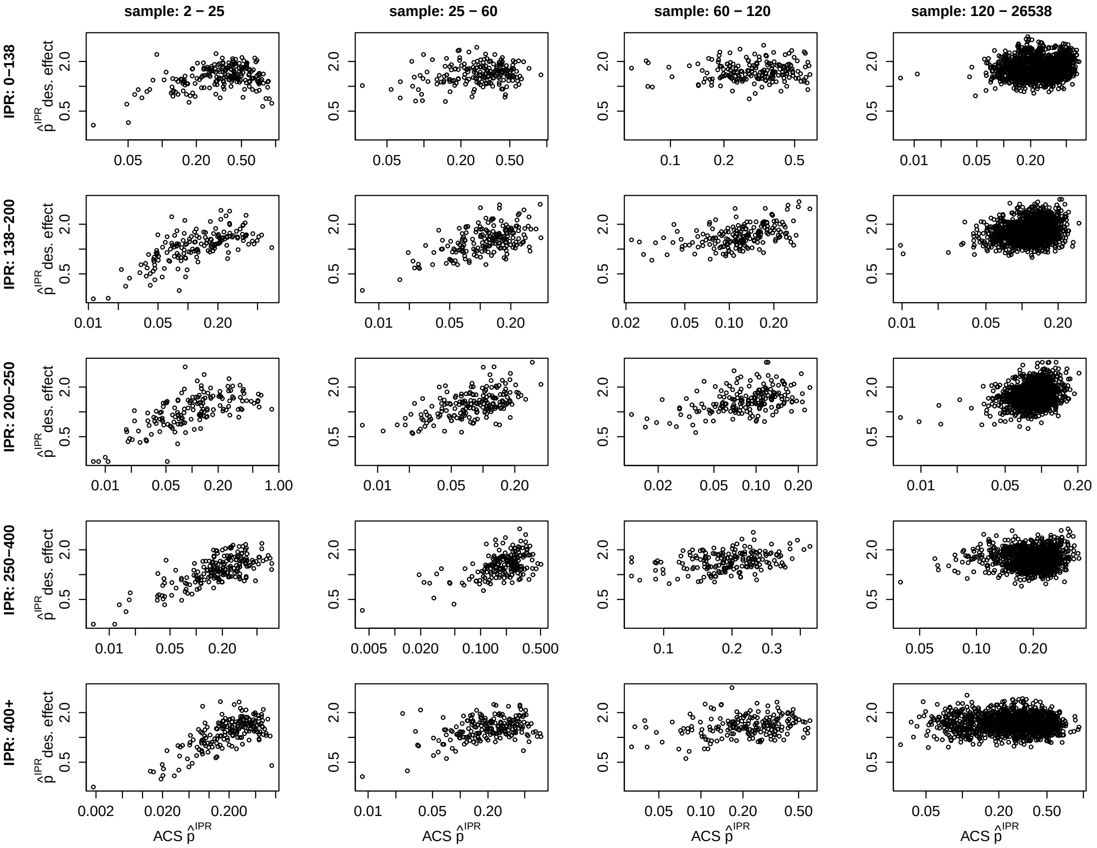

Figure 9. Plots of estimated design effects for ACS $\hat{p}^{\mathrm{IPR}}$ vs. $\hat{p}^{\mathrm{IPR}}$, by sample size group and IPR category. States. Source: ACS 2010 unpublished estimates.

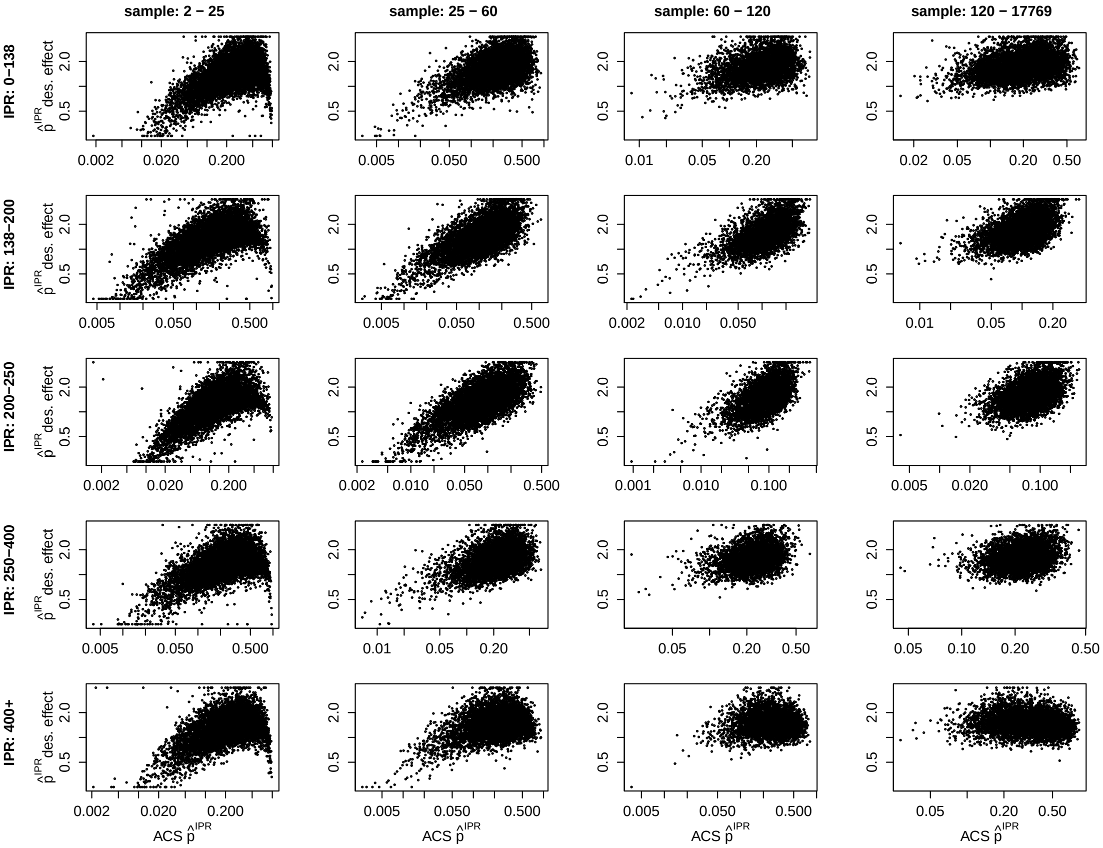

Figure 10. Plots of estimated design effects for ACS $\hat{p}^{\text{IPR}}$ vs. $\hat{p}^{\text{IPR}}$, by sample size group and IPR category. Counties. Source: ACS 2010 unpublished estimates.

6 Modeling design effects

In this section, we report results from running regression models to estimate parameters in variance function models. We model the replicate-based variance estimates. We know these are unreliable in some cases, and some plots above suggest systematic bias as well. We want to drop cases where there is substantial bias, and weight the regression to accommodate difference in reliabilities.

6.1 Weighting the regression and dropping cases

We expect reliability to be positively related to sample size. Huang and Bell (Huang and Bell (2009)) use simulations to examine properties of the variance estimation procedure "Fay's method") used in the ACS. In particular, they estimate the degrees of freedom of Fay's estimator, under the assumption that it has a χ^2_k distribution. Under that interpretation, the relative variance is two divided by the degrees of freedom, k. They argue, based on the design and number of replicates, that k should be bounded above by 78. In particular, the variance of the variance estimator does not go to zero as the sample size increases. In their simulation studies for Bernoulli populations, they found that for $p = 0.25$, k increased approximately linearly with sample size up to approximately 78. Above 78, k continued to increase, but more slowly as sample size increased (they considered sample sizes up to 780). For other values of p, the pattern is similar, but for smaller p the rate of increase of k is smaller with respect to sample size, and is close to linear for a larger range of values.

Based on the findings of Huang and Bell, for weighting purposes we assume that the inverse of the variance of the variance estimates increases linearly with sample size up to sample size 1000, and is constant after that. We thus weight observations in the regressions by the minimum of 1000 and the sample size.

Based on the findings in Figure 5, we remove cases where the unweighted count uninsured is less than five when modeling the variance of $\hat{p}^{\text{IC}}$. We assume that variances of $\hat{p}^{\text{IPR}}$ behave similarly, and remove cases in which the unweighted count in the IPR category is less than five when modeling the variance of $\hat{p}^{\text{IPR}}$. This eliminates some cases where there was a fairly clear and substantial downward bias.

In another study, Huang and Bell (Huang and Bell (2010)) conduct simulation studies to assess the bias of several variance estimators, including Fay's method used here. For Bernoulli data, they did simulations for $p = 0.08$ and $p = 0.25$. They found that Fay's method is biased downward for small samples, but that the bias is negligible for sample sizes larger than 50. For modeling the variances of $\hat{p}^{\text{IC}}$ here, we exclude cases where either the unweighted count uninsured is less than five, or the sample size (the unweighted count in the state/age/race/sex/IPR or county/age/sex/IPR category) is less than or equal to 50. For modeling the variances of $\hat{p}^{\text{IPR}}$, we exclude cases where the unweighted count in the IPR category is less than five, or the sample size (the unweighted count in the state/age/race/sex or county/age/sex group) is less than or equal to 50.

6.2 Matching the SAHIE parameterization for $\hat{p}^{\text{IC}}$

We begin modeling the design effects for $\hat{p}^{\text{IC}}$ by matching the model assumed in SAHIE. In current SAHIE modeling, we assume that

$$\text{var}(\hat{p}^{\text{IC}}_{ai}) = \lambda_0 \frac{p^{\text{IC}}_{ai}(1 - p^{\text{IC}}_{ai})}{S^{\lambda_1}_{ai}} \ . \tag{15}$$

For states, both λ_0 and λ_1 vary by two age and three IPR categories. For counties, λ_0 varies by two age and three IPR categories, but λ_1 varies only by three IPR categories.

Here, we fit the model

$$\log\left(\widehat{\text{deff}}^{\text{IC}}_{ai}\right) = \alpha + \beta \log(\text{sample size}) + \epsilon_{ai} \tag{16}$$

where α and β are allowed to vary in the same way as λ_0 and λ_1. The exponential of α corresponds to λ_0 of (15). The value of $1 - \beta$ corresponds to λ_1, because the design effect already has the sample size in its denominator.

Table 1 contains estimates of λ_0 and λ_1 for states from SAHIE 2010, and results from fitting model (16), transforming α and β to be comparable to λ_0 and λ_1. We find that the estimates of λ_1 from fitting the design effect model are significantly smaller than 1.0 as they are in SAHIE. Table 2 contains similar results from SAHIE and the design effect model for counties. In this case, the estimates of λ_1 from fitting the design effect model are much closer to 1.0 than they were for SAHIE. Correspondingly, the estimates from fitting the design effect model are larger for λ_0.

Table 1. Posterior means and standard deviations for parameters in the $\hat{p}^{\text{IC}}$ variance function from SAHIE 2010, and corresponding parameter estimates from a GVF model using direct variance estimates. Based on 5,158 observations, for states.

			SAHIE	SAHIE	GVF	GVF
parameter	age	IPR	mean	st. dev.	est.	st. err.
λ_0	0-18	0-200	1.201	0.058	1.011	0.062
	0-18	200-400	1.349	0.061	1.071	0.070
	0-18	400+	1.135	0.077	0.929	0.113
	19-64	0-200	1.189	0.067	1.318	0.043
	19-64	200-400	1.207	0.064	1.207	0.043
	19-64	400+	1.357	0.155	1.140	0.053
λ_1	0-18	0-200	0.872	0.016	0.892	0.013
	0-18	200-400	0.918	0.016	0.910	0.014
	0-18	400+	0.906	0.020	0.906	0.021
	19-64	0-200	0.911	0.017	0.983	0.009
	19-64	200-400	0.925	0.015	0.961	0.009
	19-64	400+	0.955	0.027	0.954	0.010

Source: ACS 2010 unpublished estimates, SAHIE.

Table 2. Posterior means and standard deviations for parameters in the $\hat{p}^{\text{IC}}$ variance function from SAHIE 2010, and corresponding parameter estimates from a GVF model using direct variance estimates. Based on 9,763 observations, for counties.

			SAHIE	SAHIE	GVF	GVF
parameter	age	IPR	mean	st. dev.	est.	st. err.
λ_0	0-18	0-200	1.256	0.012	1.896	0.047
	0-18	200-400	1.270	0.012	1.549	0.051
	0-18	400+	1.328	0.022	2.077	0.063
	19-64	0-200	1.035	0.008	1.387	0.046
	19-64	200-400	1.084	0.009	1.252	0.048
	19-64	400+	1.167	0.015	1.671	0.053
λ_1		0-200	0.823	0.004	0.994	0.010
		200-400	0.858	0.004	0.966	0.011
		400+	0.904	0.006	1.019	0.011

Source: ACS 2010 unpublished estimates, SAHIE.

6.3 Matching the SAHIE parameterization for $\hat{p}^{\text{IPR}}$

To match the parameterization of SAHIE in modeling the design effect for $\hat{p}^{\text{IPR}}$, We fit the model

$$\log\left(\widehat{\text{deff}}_{ai}^{\text{IPR}}\right) = \alpha + \beta \log(\text{sample size}) + \epsilon_{ai} \tag{17}$$

where α and β differ by age for states. For counties, α varies by the five ages and β varies only by two age groups.

Tables 3 and 4 contain results for states and counties respectively. In both cases, the estimate from the design effect model of λ_1 is significantly different from 1.0 for the 0 to 17 and 18 age groups for states and for 0 to 18 age group for counties, while it is nearly 1.0 for the adult age groups. Some of the SAHIE estimates of λ_0 seem unreasonably large. In those cases, the SAHIE estimate of λ_1 is larger than 1.0, indicating that the variance decreases faster than the inverse of the sample size. A possible explanation is that the estimates of λ_1 are substantially larger than one, which may have a large effect when the sample sizes are very large, tending to pull the variance estimates down. The large values of λ_1 may be to compensate, so that the variance estimates for large sample sizes fit better. At least for states, for variances of $\hat{p}^{\text{IPR}}$, it appears that fitting a variance model to the ACS variance estimates gives more reasonable results than modeling them as done in SAHIE. In the other cases, it appears that estimates from modeling the design effects confirm the ways that SAHIE estimates differ from what theory suggests about the dependence of variance on sample size. Recall that the plots, too, appear to confirm that variances decrease slower than the inverse of the sample size.

Table 3. Posterior means and standard deviations for parameters in the $\hat{p}^{\text{IPR}}$ variance function from SAHIE 2010, and corresponding parameter estimates from a GVF model using direct variance estimates. Based on 5,650 observations, for states.

parameter	age	SAHIE mean	SAHIE st. dev.	GVF est.	GVF st. err.
λ_0	0-17	4.981	0.998	1.404	0.037
	18	1.625	0.177	1.078	0.054
	19-39	3.168	0.705	1.364	0.037
	40-49	1.644	0.256	1.343	0.039
	50-64	2.011	0.331	1.374	0.035
λ_1	0-17	1.090	0.033	0.948	0.007
	18	1.022	0.028	0.947	0.011
	19-39	1.096	0.038	0.983	0.007
	40-49	0.998	0.029	0.994	0.008
	50-64	1.056	0.031	0.997	0.007

Source: ACS 2010 unpublished estimates, SAHIE.

Table 4. Posterior means and standard deviations for parameters in the $\hat{p}^{\text{IPR}}$ variance function from SAHIE 2010, and corresponding parameter estimates from a GVF model using direct variance estimates. Based on 54,406 observations, for counties.

parameter	age	SAHIE mean	SAHIE st. dev.	GVF est.	GVF st. err.
λ_0	0-17	1.963	0.056	1.625	0.021
	18	1.257	0.022	1.109	0.022
	19-39	1.384	0.029	1.430	0.013
	40-49	1.243	0.024	1.365	0.013
	50-64	1.205	0.026	1.346	0.013
λ_1	0-18	0.951	0.007	0.954	0.005
	19-64	0.938	0.005	0.988	0.003

Source: ACS 2010 unpublished estimates, SAHIE.

6.4 Dependence of design effects on the proportion

The simple random sampling variance of a proportion, p, is proportional to $p(1-p)$. Plots in Section 4 suggest that the design effects depend on p, in which case the variance is not proportional to $p(1-p)$. Here, we expand the model in (15) to include a term $p(1-p)$

$$\log\left(\widehat{\text{deff}}_{ai}\right) = \alpha + \beta \log(\text{sample size}) + \eta \log(\,\hat{p}_{ai}(1-\hat{p}_{ai})\,) + \epsilon_{ai}\ . \tag{18}$$

Here, $\hat{p}$ can be $\hat{p}^{\text{IC}}$ or $\hat{p}^{\text{IPR}}$. For consistency and ease of interpretation, we allow the coefficient, η, of $\hat{p}(1-\hat{p})$ term to vary by the same categories as the α (and hence λ_0) parameter. Note that the corresponding variance function is

$$\text{var}(\hat{p}_{ai}^{\text{IC}}) = \lambda_0 \left(p_{ai}^{\text{IC}}(1-p_{ai}^{\text{IC}})\right)^{1+\eta} \frac{1}{S_{ai}^{\lambda_1}} \tag{19}$$

where $\lambda_0 = \exp(\alpha)$ and $\lambda_1 = 1-\beta$. Under simple random sampling, α, β and η are all zero.

We initially fit the model for the design effects of $\hat{p}^{\text{IC}}$ for all observations with unweighted count uninsured larger than four and sample size larger than 50, as we had done previously. Similarly, for the design effects of $\hat{p}^{\text{IPR}}$, we used observations with unweighted count in the IPR category larger than four, and sample size larger than 50. We saw that the parameter estimates indicate η significantly larger than zero. Because these results are somewhat surprising, and because the variance estimates appear to be unreliable for small sample sizes, we fit the models again using stricter criteria for inclusion in the model fitting. For design effects for $\hat{p}^{\text{IC}}$, we required that the unweighted count uninsured be larger than 10 and that the sample size be larger than 200. For the design effects for $\hat{p}^{\text{IPR}}$, we required that the unweighted count in the IPR category be larger than 10 and the sample size larger than 200. We report results from both fits of the models.

Table 5 contains parameter estimates for the η's for states and counties for the model in (18), using the less restrictive sample size criteria. Table 6 contains parameter estimates when the model was fit using the more restrictive criteria, according to which we only include observations with an unweighted uninsured count larger than 10 and sample size larger than 200. For these latter results, we only allowed parameters to vary by two ages, because there were many fewer records that meet the criteria.

Table 7 contains parameter estimates for the model in (18) but for the estimates, $\hat{p}^{\text{IPR}}$, of proportions in the IPR categories. Table 8 contains parameter estimates for the variance of $\hat{p}^{\text{IPR}}$, but using the stricter sample size criteria.

When fit using the larger number of observations, the estimate of the η parameter was positive and significantly different from zero for all categories, for both $\hat{p}^{\text{IC}}$ and $\hat{p}^{\text{IPR}}$. These results suggest that the variance of $\hat{p}$ is proportional to $p(1-p)$ raised to a power of between about 1.1 and 1.4.

When we fit the model using only observations meeting stricter criteria, the estimates of η become closer to zero. For $\hat{p}^{\text{IC}}$, the estimates remain positive and significantly different from zero for children, but effectively zero for adults. For $\hat{p}^{\text{IPR}}$, the estimates are still positive and significantly different from zero for both children and adults, but not large. The estimates for adults are closer to zero than for children, as with the parameters for variances of $\hat{p}^{\text{IC}}$.

The results suggest that the variances decrease more rapidly as p approaches 0 or 1 than in simple random sampling. However, we found that the estimates of the parameter η in (18) are sensitive to the choice of observations to include in the model fitting. Because of this, it seems unclear whether the difference from zero of estimates of η is a fact about the variances or an artifact of the variance estimation method.

6.5 The effect of collection mode

One factor that we expect to affect ACS sampling variances is the prevalence of different modes of collection. ACS uses three modes of data collection: mail, telephone, and Computer Assisted Personal Interviewing (CAPI) (U.S. Census Bureau U.S. Census Bureau (2009), Chapter 4). Addresses initially selected for the sample, but from which neither a mail questionnaire nor a telephone interview has been completed, are sampled for CAPI. The CAPI sampling rate is 67 percent for unmailable addresses (incomplete addresses and those that refer to a post office box) and some addresses in remote Alaska. The CAPI sampling rates range from 33 percent to 50 percent for other addresses, based on the predicted levels of completed interviews prior to CAPI for the tract.

We expect sampling variances to be different when there is a larger proportion of CAPI cases both because the average sampling rate is smaller, and because there will be greater variation in the sampling weights because a greater proportion of cases will have the larger CAPI weights. The design effect models we fit

Table 5. Parameter estimates for η in the design effect model (18), for $\hat{p}^{\text{IC}}$. Based on 5,158 observations for states, and 9,763 observations for counties.

geography	age	IPR	est.	st. err.	Pr $> \|t\|$
state	0-18	0-200	0.199	0.020	$< .0001$
	0-18	200-400	0.269	0.022	$< .0001$
	0-18	400+	0.291	0.043	$< .0001$
	19-64	0-200	0.046	0.027	0.091
	19-64	200-400	0.152	0.020	$< .0001$
	19-64	400+	0.234	0.022	$< .0001$
county	0-18	0-200	0.429	0.018	$< .0001$
	0-18	200-400	0.366	0.024	$< .0001$
	0-18	400+	0.428	0.041	$< .0001$
	19-64	0-200	0.100	0.023	$< .0001$
	19-64	200-400	0.194	0.013	$< .0001$
	19-64	400+	0.293	0.016	$< .0001$

Source: ACS 2010 unpublished estimates.

Table 6. Parameter estimates for η in the design effect model (18), for $\hat{p}^{\text{IC}}$. Here, only used records with an unweighted count of uninsured larger than 10 and sample size larger than 200. Based on 2,903 observations for states, and 2,179 observations for counties.

geography	age	est.	st. err.	Pr $> \|t\|$
state	0-18	0.156	0.015	$< .0001$
	19-64	-0.005	0.010	0.591
county	0-18	0.169	0.021	$< .0001$
	19-64	0.017	0.010	0.079

Source: ACS 2010 unpublished estimates.

Table 7. Parameter estimates for η in the design effect model of (18), for $\hat{p}^{\text{IPR}}$. Based on 5,650 observations for states, and 54,406 observations for counties.

geography	age	est.	st. err.	Pr $> \|t\|$
state	0-17	0.242	0.018	$< .0001$
	18	0.180	0.015	$< .0001$
	19-39	0.172	0.020	$< .0001$
	40-49	0.121	0.017	$< .0001$
	50-64	0.126	0.016	$< .0001$
county	0-17	0.329	0.006	$< .0001$
	18	0.155	0.021	$< .0001$
	19-39	0.217	0.006	$< .0001$
	40-49	0.146	0.006	$< .0001$
	50-64	0.132	0.005	$< .0001$

Source: ACS 2010 unpublished estimates.

Table 8. Parameter estimates for η in the design effect model of (18), for $\hat{p}^{\text{IPR}}$. Here, only used observations with unweighted count in the IPR category larger than 10 and sample size larger than 200. Based on 3,807 observations for states, and 15,772 observations for counties.

geography	age	est.	st. err.	Pr $> \|t\|$
state	0-18	0.070	0.015	$< .0001$
	19-64	0.031	0.011	0.007
county	0-18	0.059	0.010	$< .0001$
	19-64	0.024	0.005	$< .0001$

Source: ACS 2010 unpublished estimates.

already take sample size into account. But the difference in weights for the CAPI sample may have an additional effect on sampling variances.

To assess the value of the proportion of CAPI cases in predicting sampling variances, we fit the model as in (18), but add the proportion of CAPI cases, $PCAPI$, to the model. For the variances of $\hat{p}^{\text{IC}}$, the resulting model is

$$\begin{aligned}\log\left(\widehat{\text{deff}}_{ai}\right) &= \alpha + \beta\log(S_{ai}) + \eta\log(\,\hat{p}^{\text{IC}}_{ai}(1-\hat{p}^{\text{IC}}_{ai})\,)\\ &\quad + \gamma PCAPI_{ai} + \epsilon_{ai}\ . \end{aligned} \tag{20}$$

Table 9 contains estimates of the parameter γ when we fit the model in (20), weighting observations as before. We use the less strict criteria of a sample size larger than 50 and unweighted count uninsured larger than four. We allowed the parameter γ to vary in the same way as α and η. Table 10 contain estimates when using the stricter criteria of a sample size larger than 200 and unweighted count uninsured larger than 10. As before, we allowed the parameter to vary by just two ages because of the smaller number of observations used.

When we used the less strict criteria, for states in the 0 to 18 age group, the estimated $\hat{\gamma}$ of coefficient of the CAPI rate, is significantly different from zero for one of the three IPR groupings, and is positive in that case. For the 19 to 64 age groups, $\hat{\gamma}$ is significantly different from zero for two of the three IPR groupings, and is negative in those cases. For counties, using the less strict criteria, the coefficient estimate, $\hat{\gamma}$, is significantly different from zero for two of three IPR groupings in both the 0 to 18 and 19 to 64 age groupings. In those cases, the $\hat{\gamma}$ is always negative. The negative parameter estimates indicate that the design effect estimates tend to be smaller when the CAPI rate is larger, which seems surprising.

These results are confirmed when we use only observations meeting the stricter criteria. In that case, for states, $\hat{\gamma}$ is not significantly different from zero for the age group 0 to 18, and is negative for age grouping 19 to 64. For counties, γ is negative for both age groupings.

Table 9. Parameter estimates for γ in the design effect model of (20), for $\hat{p}^{\text{IC}}$. Model fit using observations with unweighted count uninsured larger than 4 and sample size larger than 50. Based on 5,158 observations for states, and 9,763 observations for counties.

geography	age	IPR	est.	st. err.	Pr > $\lvert t\rvert$
state	0-18	0-200	0.008	0.076	0.9191
	0-18	200-400	0.233	0.093	0.0119
	0-18	400+	0.147	0.238	0.5369
	19-64	0-200	-0.065	0.046	0.1603
	19-64	200-400	-0.223	0.059	0.0001
	19-64	400+	-0.260	0.119	0.0286
county	0-18	0-200	-0.325	0.064	< 0.0001
	0-18	200-400	-0.468	0.117	0.0001
	0-18	400+	0.097	0.451	0.8297
	19-64	0-200	0.023	0.045	0.6088
	19-64	200-400	-0.409	0.062	< 0.0001
	19-64	400+	-0.299	0.118	0.0111

Source: ACS 2010 unpublished estimates.

Table 10. Parameter estimates for γ in the design effect model of (20), for $\hat{p}^{\text{IC}}$. Here, the model was fit using only observations with unweighted count uninsured larger than 10 and sample size larger than 200. Based on 2,903 observations for states, and 2,179 observations for counties.

geography	age	est.	st. err.	Pr > $\|t\|$
state	0-18	-0.066	0.069	0.3409
	19-64	-0.297	0.051	< .0001
county	0-18	-0.533	0.097	< .0001
	19-64	-0.276	0.075	0.0002

Source: ACS 2010 unpublished estimates.

7 Summary and questions raised

In this paper, we considered using direct estimates of ACS sampling variances in developing models to predict ACS sampling variances. This is a different approach from that currently used in SAHIE, in which direct sampling variance estimates are not used, and instead a parametric function for sampling variances is fit jointly with the rest of the unknowns in SAHIE estimation.

We note here some findings from the research, and questions raised.

- **Some confirmation of SAHIE parameter estimates.** One reason to consider such a model is that some parameter estimates from SAHIE appear different than theory would suggest. This possibly indicates model misspecification in SAHIE. In particular, in several cases, SAHIE parameter estimates imply that sampling variances decrease with sample size at a rate slower than the inverse of the sample size. However, exploratory analysis of design effects based on direct variance estimates appear to confirm that sampling variances do decrease at a rate slower than the inverse of the sampling variance. Fitting a model to the direct variance estimates appears to confirm this, although the parameter estimates from modeling the direct variance estimates imply that sampling variances decrease at a rate closer to the inverse of the sampling variance than do SAHIE estimates. These results raise some questions: Is there sampling theory that could explain that variances decrease at a rate slower than the inverse of the sample size? If not, is there something about the variance estimation procedure that explains the appearance that the variances decrease at a rate slower than the inverse of the sample size?
- **An unexpected dependence on the term $p(1-p)$.** In SAHIE, we model ACS estimates of two proportions. A notable finding from fitting a model to the direct variance estimates is that the variances of either of the proportions, p, are not proportional to $p(1-p)$. Instead, the variances appear to be proportional to $p(1-p)$ raised to a power larger than one. This raises questions: Is there a sampling model that would explain that the variance of the estimate of a proportion, p, decreases as p approaches zero or one more rapidly than $p(1-p)$? If not, is there an explanation of why the variance estimation procedure results in estimates of variances that approach zero and one more rapidly than theory predicts?
- **Biases in variance estimates and how to handle them.** There are issues evident here for an approach that models sampling variance estimates to predict sampling variances. Previous research (e.g. Huang and Bell (2010)) had shown bias in the direct variance estimates for small sample sizes. The research in the present paper (in Figures 4, 5, 9, 10) suggests that the bias for the estimate of a proportion is more pronounced as the proportion approaches zero or one. These results suggest dropping from the model observations with small sample sizes or small counts related to the proportion of interest. We fit models restricting the observations to those with sample sizes at which Huang and Bell noted negligible bias. But when we refit the models, using stricter criteria on sample sizes, we get different estimates. This raises a question: If we are to fit variance models using only observations fitting some criteria, what criteria should we use? Further, if those criteria involve omitting observations with small sample sizes, then we would likely assume that the same model can be extrapolated from large sample sizes to small sample sizes. Can that assumption be justified or verified? Note that this latter question applies to the current SAHIE procedure under which sampling variances are modeled.

But while it is true that SAHIE, too, assumes a single model for large and small sample sizes, SAHIE modeling at least allows all observations to be relevant to the modeling.

- **General issue raised by some of the previous questions.** When using estimates of sampling variances, what are artifacts of the estimation procedure and what are truly features of the sampling variances?

References

Huang, E. and W. Bell (2009). A simulation study of the distribution of Fay's successive difference replication variance estimator. In *Proceedings of the American Statistical Association, Section on Survey Research Methods*, pp. 5294–5308.

Huang, E. and W. Bell (2010). Further simulation results on the distribution of some survey variance estimators. In *Proceedings of the American Statistical Association, Section on Survey Research Methods*, pp. 3877–3889.

Judkins, D. (1990). Fay's method for variance estimation. *Journal of Official Statistics 6*(3), 223–239.

U.S. Census Bureau (2009). *Design and Methodology: American Community Survey.* Washington, DC: U.S. Government Printing Office.
`http://www.census.gov/acs/www/methodology/methodology_main`.